The Annals of Probability
2011, Vol. 39, No. 3, 1137–1160
DOI: 10.1214/10-AOP576

NONPARAMETRIC SEQUENTIAL PREDICTION FOR STATIONARY PROCESSES

By Gusztáv Morvai[1] and Benjamin Weiss

MTA-BME Stochastics Research Group and Hebrew University of Jerusalem

We study the problem of finding an universal estimation scheme $h_n : \mathbb{R}^n \to \mathbb{R}$, $n = 1, 2, \ldots$ which will satisfy

$$\lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{t} |h_i(X_0, X_1, \ldots, X_{i-1}) - E(X_i | X_0, X_1, \ldots, X_{i-1})|^p = 0 \qquad \text{a.s.}$$

for all real valued stationary and ergodic processes that are in L^p. We will construct a single such scheme for all $1 < p \le \infty$, and show that for $p = 1$ mere integrability does not suffice but $L \log^+ L$ does.

1. Introduction. The problem of sequentially predicting the next value X_n of a stationary process after observing the initial values X_i for $0 \le i < n$ is one of the central problems in probability and statistics. Usually, one bases the prediction on the conditional expectation $E(X_n | X_0^{n-1})$ where we write for brevity $X_0^{n-1} = \{X_0, X_1, \ldots, X_{n-1}\}$. However, when one does not know the distribution of the process one is faced with the problem of estimating the conditional expectation from a single sample of length n. It was shown long ago by Bailey [5] (cf. also Ryabko [30] and Györfi Morvai and Yakowitz [10]) that even for binary processes no universal scheme $h_n(X_0^{n-1})$ exists which will almost surely satisfy $\lim_{n\to\infty}(h_n(X_0^{n-1}) - E(X_n | X_0^{n-1})) = 0$. This is in contrast to the backward estimation problem where one is trying to estimate $E(X_0 | X_{-\infty}^{-1})$ based on the successive observations of $X_{-\infty}^{-1}$. Here, it was Ornstein [29] who constructed the first such universal estimator for finite valued processes. This was generalized to bounded processes by Algoet [1], Morvai [16] and Morvai Yakowitz and Györfi [18]. For unbounded processes,

Received November 2007; revised May 2010.
[1]Supported by the Bolyai János Research Scholarship and OTKA Grant K75143.
AMS 2000 subject classifications. 60G45, 60G25, 62G05.
Key words and phrases. Nonparametric predicton, stationary processes.

several universal estimators were constructed (see Algoet [3] and Györfi et al. [9]).

Returning to our original problem of sequential prediction it was already observed by Bailey that backward schemes could be used for the sequential prediction problem successfully in the sense that that the error tends to zero in the Cesáro mean. To establish this, he applied a generalized ergodic theorem which requires some technical hypotheses which were satisfied in his case.

Over the years some authors have extended this work, namely of adapting backward schemes to sequential prediction, but only for bounded processes (see Algoet [1, 3], Morvai [16], Morvai Yakowitz and Györfi [18] and Györfi et al. [9]).

Another approach to the sequential prediction used a weighted average of expert schemes, and with these results were extended to the general unbounded case by Nobel [28] and Ottucsak [12] (see also the survey of Feder and Merhav [8]). However, none of these results were optimal in the sense that moment conditions higher than necessary were assumed. It is our purpose to obtain these optimal conditions and to show why they are necessary. We consider the following problem for $1 \le p \le \infty$. Does there exist a scheme $h_n(X_0^{n-1})$ which will satisfy

$$\lim_{t\to\infty} \frac{1}{t}\sum_{i=1}^{t} |h_i(X_0^{i-1}) - E(X_i|X_0^{i-1})|^p = 0 \qquad \text{a.s.}$$

for all real valued stationary and ergodic processes that are in L^p. The only case that has been solved completely is when p is infinity. Even the recent schemes Nobel [28] and Györfi and Ottucsak [12] put a higher moment condition on the process than is manifestly required. Our main result is that the basic scheme first introduced by the first author in his thesis can be adapted to give a scheme which will answer our problem positively for all $1 < p$. For $p = 1$, we shall show that stronger hypothesis is necessary, as is usually the case, and we will establish the convergence under the hypothesis that $X_0 \in L\log^+ L$.

In the third section, we will show how this hypothesis cannot be weakened to $X_0 \in L^1$. Our construction will be based on one of the simplest ergodic transformation, the adding machine, and illustrates the richness of behavior that is possible for processes that are almost periodic (in the sense of Besicovich).

As soon as one knows that the errors converge to zero in Cesáro mean, it follows that there is a set of density one of time moments along which the errors converge to zero. However, in general one does not know what this sequence is. In the framework of estimation, schemes adapted to a sequence of stopping times (see [19–21, 23–26]) one may ask can one find a sequence

of stopping times with density one along which the errors of a universal sequential prediction scheme will tend to zero. We have been unable to do this in general and regard it as an important open problem. Finally, we refer the interested reader to some other papers which are relevant to this line of research [2, 11, 17, 27, 34].

Some technical probabilistic results have been relegated to the Appendix, they are of a classical nature and may be known, but we were unable to find references.

2. The main result. Let $\mathbf{X} = \{X_n\}$ denote a real-valued doubly infinite stationary ergodic time series. Let

$$X_i^j = (X_i, X_{i+1}, \dots, X_j)$$

be notation for a data segment, where i may be minus infinity. Let

$$\mathbf{X}^- = X_{-\infty}^{-1}.$$

Let G_k denote the quantizer

$$G_k(x) = \begin{cases} 0, & \text{if } -2^{-k} < x < 2^{-k}, \\ -i2^{-k}, & \text{if } -(i+1)2^{-k} < x \le -i2^{-k} \text{ for some } i = 1, 2, \dots, \\ i2^{-k}, & \text{if } i2^{-k} \le x < (i+1)2^{-k}. \end{cases}$$

Define the sequences λ_{k-1} and τ_k recursively $(k = 1, 2, \dots)$. Put $\lambda_0 = 1$ and let τ_k be the time between the occurrence of the pattern

$$B(k) = (G_k(X_{-\lambda_{k-1}}), \dots, G_k(X_{-1})) = G_k(X_{-\lambda_{k-1}}^{-1})$$

at time -1 and the last occurrence of the same pattern prior to time -1. More precisely, let

$$\tau_k = \min\{t > 0 : G_k(X_{-\lambda_{k-1}-t}^{-1-t}) = G_k(X_{-\lambda_{k-1}}^{-1})\}.$$

Put

$$\lambda_k = \tau_k + \lambda_{k-1}.$$

Define

$$R_k = \frac{1}{k} \sum_{1 \le j \le k} X_{-\tau_j}. \tag{2.1}$$

To obtain a fixed sample size $t > 0$ version, let κ_t be the maximum of integers k for which $\lambda_k \le t$. For $t > 0$, put

$$\hat{R}_{-t} = \frac{1}{\kappa_t} \sum_{1 \le j \le \kappa_t} X_{-\tau_j}. \tag{2.2}$$

Motivated by Bailey [5], for $t > 0$ consider the estimator

$$\hat{R}_t(\omega) = \hat{R}_{-t}(T^t\omega),$$

which is defined in terms of $(X_0, \dots, X_{t-1})$ in the same way as $\hat{R}_{-t}(\omega)$ was defined in terms of $(X_{-t}, \dots, X_{-1})$. (T denotes the left shift operator.) The estimator $\hat{R}_t$ may be viewed as an online predictor of X_t. This predictor has special significance not only because of potential applications, but additionally because Bailey [5] proved that it is impossible to construct estimators $\hat{R}_t$ such that always $\hat{R}_t - E(X_t|X_0^{t-1}) \to 0$ almost surely.

THEOREM 1. *Let $\{X_n\}$ be stationary and ergodic. Assume that*

$$E(|X_0|\log^+(|X_0|)) < \infty.$$

Then

$$\lim_{t\to\infty} \frac{1}{t}\sum_{i=1}^{t} |\hat{R}_i - E(X_i|X_0^{i-1})| = 0 \qquad a.s. \tag{2.3}$$

and

$$\lim_{t\to\infty} \frac{1}{t}\sum_{i=1}^{t} |\hat{R}_i - X_i| = E(|E(X_0|X_{-\infty}^{-1}) - X_0|) \qquad a.s. \tag{2.4}$$

Furthermore, if for some $1 < p < \infty$, $E(|X_0|^p) < \infty$, then

$$\lim_{t\to\infty} \frac{1}{t}\sum_{i=1}^{t} |\hat{R}_i - E(X_i|X_0^{i-1})|^p = 0 \qquad a.s. \tag{2.5}$$

and

$$\lim_{t\to\infty} \frac{1}{t}\sum_{i=1}^{t} |\hat{R}_i - X_i|^p = E(|E(X_0|X_{-\infty}^{-1}) - X_0|^p) \qquad a.s. \tag{2.6}$$

PROOF. The proof will follow the same pattern in all four cases. We will verify that the backward estimator scheme converges almost surely and we will see that the sequence of errors is dominated by an integrable function. This allows us to conclude from the generalized ergodic theorem of Maker (rediscovered by Breiman, cf. Theorem 1 in Maker [15] or Theorem 12 in Algoet [2]) that the forward scheme converges in Cesaro mean. For the first case, we will carry this out in full detail, for the others we will just check the requisite properties for the backward scheme. First, consider

$$R_k = \frac{1}{k}\sum_{1\le j\le k} [X_{-\tau_j} - G_j(X_{-\tau_j})]$$

$$
\begin{aligned}
&+\frac{1}{k}\sum_{1\le j\le k}[G_j(X_{-\tau_j})-E(G_j(X_{-\tau_j})|G_{j-1}(X^{-1}_{-\lambda_{j-1}}))]\\
&+\frac{1}{k}\sum_{1\le j\le k}[E(G_j(X_{-\tau_j})|G_{j-1}(X^{-1}_{-\lambda_{j-1}}))-E(X_{-\tau_j}|G_{j-1}(X^{-1}_{-\lambda_{j-1}}))]\\
&+\frac{1}{k}\sum_{1\le j\le k}[E(X_{-\tau_j}|G_{j-1}(X^{-1}_{-\lambda_{j-1}}))-E(X_0|G_{j-1}(X^{-1}_{-\lambda_{j-1}}))]\\
&+\frac{1}{k}\sum_{1\le j\le k}E(X_0|G_{j-1}(X^{-1}_{-\lambda_{j-1}}))\\
&=A_k+B_k+C_k+D_k+E_k.
\end{aligned}
$$

Obviously,

$$|A_k|+|C_k|\le\frac{2}{k}\sum_{1\le j\le k}2^{-j}\le\frac{2}{k}\to 0.$$

Now we will deal with D_k. By Lemma 1, in Morvai, Yakowitz and Györfi [18],

$$P(X_{-\tau_j}\in C|G_{j-1}(X^{-1}_{-\lambda_{j-1}}))=P(X_0\in C|G_{j-1}(X^{-1}_{-\lambda_{j-1}})).$$

Using this, we get that $D_k=0$.

Assume that $E(|X_0|\log^+(|X_0|))<\infty$. Toward mastering B_k, one observes that $\{X_{-\tau_j}\}$ are identically distributed by Lemma 1 in Morvai, Yakowitz and Györfi [18] and B_k is an average of martingale differences. By Proposition 1 in the Appendix, $|B_k|\to 0$ almost surely and $E(\sup_{1\le k}|B_k|)<\infty$.

Now we deal with the last term E_k. By assumption,

$$\sigma(G_j(X^{-1}_{-\lambda_j}))\uparrow\sigma(\mathbf{X}^-).$$

Consequently by the a.s. martingale convergence theorem, we have that

$$E(X_0|G_j(X^{-1}_{-\lambda_j}))\to E(X_0|\mathbf{X}^-)\qquad\text{a.s.},$$

and thus

$$E_k\to E(X_0|\mathbf{X}^-)\qquad\text{a.s.}$$

Furthermore, by Doob's inequality, cf. Theorem 1 on page 464, Section 3, Chapter VII in Shiryayev [32], $E(\sup_{1\le k}|E_k|)\le E(\sup_{1\le j}|E(X_0|G_j(X^{-1}_{-\lambda_j}))|)<\infty$.

We have so far proved that

$$R_k\to E(X_0|\mathbf{X}^-)\qquad\text{almost surely}$$

and

$$E\Big(\sup_{1\le k}|R_k|\Big)<\infty.$$

This in turn implies that

$$\lim_{t\to\infty}\hat{R}_{-t}=E(X_0|\mathbf{X}^-)\qquad\text{almost surely}$$

and

$$E\Big(\sup_{1\le t}|\hat{R}_{-t}|\Big)<\infty.$$

Now since $E(X_0|X_{-t}^{-1})\to E(X_0|\mathbf{X}^-)$ almost surely,

$$\lim_{t\to\infty}|\hat{R}_{-t}-E(X_0|X_{-t}^{-1})|=0\qquad\text{almost surely}$$

and by Doob's inequality,

$$\begin{aligned}E\Big(\sup_{1\le t}|\hat{R}_{-t}-E(X_0|X_{-t}^{-1})|\Big)&\le E\Big(\sup_{1\le t}|\hat{R}_{-t}|\Big)\\&\qquad+E\Big(\sup_{1\le t}|E(X_0|X_{-t}^{-1})|\Big)\\&<\infty.\end{aligned}$$

Now, apply the generalized ergodic theorem to conclude that

$$\begin{aligned}\lim_{t\to\infty}\frac{1}{t}\sum_{i=1}^{t}(|\hat{R}_{-i}-E(X_0|X_{-i}^{-1})|(T^i\omega))&=\lim_{t\to\infty}\frac{1}{t}\sum_{i=1}^{t}|\hat{R}_i-E(X_i|X_0^{i-1})|\\&=0\qquad\text{a.s.}\end{aligned}$$

and the proof of (2.3) is complete. Similarly,

$$\lim_{t\to\infty}|\hat{R}_{-t}-X_0|=|E(X_0|X_{-\infty}^{-1})-X_0|\qquad\text{almost surely}$$

and

$$E\Big(\sup_{1\le t}|\hat{R}_{-t}-X_0|\Big)\le E\Big(\sup_{1\le t}|\hat{R}_{-t}|\Big)+E(|X_0|)<\infty$$

and the generalized ergodic theorem gives

$$\begin{aligned}\lim_{t\to\infty}\frac{1}{t}\sum_{i=1}^{t}(|\hat{R}_{-i}-X_0|(T^i\omega))&=\lim_{t\to\infty}\frac{1}{t}\sum_{i=1}^{t}|\hat{R}_i-X_i|\\&=E(|E(X_0|X_{-\infty}^{-1})-X_0|)\qquad\text{a.s.}\end{aligned}$$

and the proof of (2.4) is complete.

Now, we assume that for some $1 < p < \infty$, $E(|X_0|^p) < \infty$, and we prove (2.5). Observe that

$$|R_k|^p \le 3^p\left[\left(\frac{2}{k}\right)^p + |B_k|^p + |E_k|^p\right]$$

and since by Proposition 2 in the Appendix $|B_k| \to 0$ almost surely and $E(\sup_{1\le k}|B_k|^p) < \infty$ and by Doob's inequality, $E(\sup_{1\le k}|E_k|^p) < \infty$ and $E_k \to E(X_0|\mathbf{X}^-)$ almost surely (for the same reason as before).

We have so far proved that

$$R_k \to E(X_0|\mathbf{X}^-) \qquad \text{almost surely}$$

and

$$E\Big(\sup_{1\le k}|R_k|^p\Big) < \infty.$$

This in turn implies that

$$\lim_{t\to\infty} \hat{R}_{-t} = E(X_0|\mathbf{X}^-) \qquad \text{almost surely}$$

and

$$E\Big(\sup_{1\le t}|\hat{R}_{-t}|^p\Big) < \infty.$$

Now since $E(X_0|X_{-t}^{-1}) \to E(X_0|\mathbf{X}^-)$ almost surely,

$$\lim_{t\to\infty}|\hat{R}_{-t} - E(X_0|X_{-t}^{-1})|^p = 0 \qquad \text{almost surely}$$

and by Doob's inequality,

$$\begin{aligned}
E\Big(\sup_{1\le t}|\hat{R}_{-t} - E(X_0|X_{-t}^{-1})|^p\Big) &\le 2^p E\Big(\sup_{1\le t}|\hat{R}_{-t}|^p\Big) \\
&\quad + 2^p E\Big(\sup_{1\le t}|E(X_0|X_{-t}^{-1})|^p\Big) \\
&< \infty.
\end{aligned}$$

By Maker's (or Breiman's) generalized ergodic theorem (cf. Theorem 1 in Maker [15] or Theorem 12 in Algoet [2]) one gets (2.5). Similarly,

$$\lim_{t\to\infty}|\hat{R}_{-t} - X_0|^p = |E(X_0|X_{-\infty}^{-1}) - X_0|^p \qquad \text{almost surely}$$

and

$$E\Big(\sup_{1\le t}|\hat{R}_{-t} - X_0|^p\Big) \le 2^p E\Big(\sup_{1\le t}|\hat{R}_{-t}|^p\Big) + 2^p E(|X_0|^p) < \infty.$$

Now, apply Maker's (or Breiman's) generalized ergodic theorem to prove (2.6). The proof of Theorem 1 is complete. $\square$

REMARK 1. We are indebted to the referee for the following remark. Using the notion of Bochner integrability of strongly measurable functions with values in c_0 and the extension of Birkhoff's ergodic theorem to Banach space valued functions (see Krengel [14], page 167), one can give an easy proof of Maker's theorem. The key condition now becomes the fact that the norm of the sequence $\{f - f_k\}$ in c_0 is integrable, and then the convergence in the norm of c_0 allows one to deduce the convergence of the diagonal sequence which is what appears in Maker's theorem.

3. Integrability alone is not enough. In Theorem 1 for the Cesáro convergence in the L^1 norm, we assumed that X_0 was not merely in L^1 but in $L^1 \log^+ L$. In this section, we shall show that some additional condition is really necessary. We will first give an example to show that the maximal function of the conditional expectations $\sup_{1\le n}|E(X_0|X_{-n}^{-1})|$ may be non-integrable for an integrable process. We shall do so in an indirect fashion by showing that the the estimate $E(X_n|X_0^{n-1})$ for $E(X_n|X_{-\infty}^{n-1})$ does not converge in Cesáro mean to zero. This means that even though we are may be in the distant future the information of the prehistory can make a serious difference. This example serves as a model for the main result of the section where we show that for any estimation scheme for $E(X_n|X_0^{n-1})$ which converges almost surely in Cesáro mean for all bounded processes there will be some ergodic integrable process where it fails to converge. Indeed the processes that we need to consider are countably valued and in fact are zero entropy and finitarily Markovian (see below for a definition), a generalization of finite order Markov chains.

First, let us fix the notation. Let $\{X_n\}_{n=-\infty}^{\infty}$ be a stationary and ergodic time series taking values from finite or countable alphabet $\mathcal{X}$. (Note that all stationary time series $\{X_n\}_{n=0}^{\infty}$ can be thought to be a two sided time series, that is, $\{X_n\}_{n=-\infty}^{\infty}$.)

DEFINITION 1. The stationary time series $\{X_n\}$ is said to be finitarily Markovian if almost surely the sequence of the conditional distributions $\mathcal{L}$ $(X_1|X_{-k}^0)$ is constant for large k (it is random how large k should be).

This class includes of course all finite order Markov chains but also many other processes such as the finitarily determined processes of Kalikow, Katznelson and Weiss [13], which serve to represent all isomorphism classes of zero entropy processes.

For some concrete examples that are not Markovian, consider the following example.

EXAMPLE 1. Let $\{M_n\}$ be any stationary and ergodic first order Markov chain with finite or countably infinite state space S. Let $s \in S$ be an arbitrary

state with $P(M_1 = s) > 0$. Now let $X_n = I_{\{M_n = s\}}$. By Shields [31], Chapter I.2.c.1, the binary time series $\{X_n\}$ is stationary and ergodic. It is also finitarily Markovian. Indeed, the conditional probability $P(X_1 = 1 | X_{-\infty}^0)$ does not depend on values beyond the first (going backward) occurrence of one in $X_{-\infty}^0$ which identifies the first (going backward) occurrence of state s in the Markov chain $\{M_n\}$. The resulting time series $\{X_n\}$ is not a Markov chain of any order in general.

We note that Morvai and Weiss [22] proved that there is no classification rule for discriminating the class of finitarily Markovian processes from other ergodic processes. For more about estimation for finitarily Markovian processes, see Morvai and Weiss [23, 24, 26].

THEOREM 2. *Let $\mathcal{X} = \{0, 10^{-k}, \frac{2^k}{3^m}, k = 1, 2, \ldots, m = 1, 2, \ldots\}$. There exists a stationary and ergodic finitarily Markovian time series $\{X_n\}$ taking values from $\mathcal{X}$ such that $E|X_0| < \infty$ and*

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} |E(X_n | X_0^{n-1}) - E(X_n | X_{-\infty}^{n-1})| = \infty$$

almost surely. Therefore,

$$E\Big(\sup_{1\le n} |E(X_0 | X_{-n}^{-1})|\Big) = \infty.$$

PROOF. Let Ω be the one sided sequence space over $\{0, 1\}$. Let $\omega = (\omega_1, \omega_2, \ldots) \in \Omega$. Define the transformation $T : \Omega \to \Omega$ as follows:

$$(T\omega)_i = \begin{cases} 0, & \text{if } \omega_j = 1 \text{ for all } j \le i, \\ 1, & \text{if } \omega_i = 0 \text{ and for all } j < i : \omega_j = 1, \\ \omega_i, & \text{otherwise.} \end{cases}$$

Consider the product measure $P = \Pi_{i=1}^{\infty} \{1/2, 1/2\}$ on Ω which is preserved by T. It is well known (cf. Aaronson [4], page 25) that (Ω, P, T) is an ergodic process, called the adding machine or dyadic odometer. The process will be defined by a function $f : \Omega \to \mathbb{R}$ as $X_n(\omega) = f(T^n \omega)$. Let $l_3 < \cdots < l_{k-1} < l_k \to \infty$. Define $a_k = a$ and $b_k = b$ when $k = 2^a + b$ where $1 \le b \le 2^a$. Define

$$C_k = \{\omega : \omega_i = 1 \text{ for } 1 \le i < l_k, \omega_{l_k} = 0\},$$

clearly $P(C_k) = 2^{-l_k}$. Let

$$D_k = \{\omega : \omega_i = 1 \text{ for } 1 \le i < l_k - a_k, \omega_{l_k - a_k} = 0, \omega_i = 1 \text{ for } l_k - a_k < i < l_k\}$$

and

$$E_k = \bigcup_{i=0}^{2^{l_k - a_k - 1} - 1} T^{-i} D_k.$$

Notice that

$$E_k = \{\omega : \omega_{l_k - a_k} = 0, \omega_j = 1, \text{ for all } l_k - a_k < j < l_k\}.$$

It is clear that if the l_k's are chosen large enough so that for all $k' > k$ $l_k < l_{k'} - 2a_{k'}$:

- the family C_k, D_l $k, l \geq 3$ consists of disjoint sets,
- the intervals $[l_k - a_k, l_k - 1]$ are also disjoint and therefore the sets E_k are independent.

The signaling function u is defined by

$$u(\omega) = \sum_{k=3}^{\infty} 10^{-k} I_{D_k}(\omega)$$

and the main contributor to f will be

$$v(\omega) = \sum_{k=3}^{\infty} \frac{2^{l_k}}{3^{a_k}} I_{C_k}(\omega).$$

Clearly,

$$\begin{aligned} E(v(\omega)) &= \sum_{k=3}^{\infty} \frac{2^{l_k}}{3^{a_k}} P(C_k) = \sum_{k=3}^{\infty} \frac{1}{3^{a_k}} = \sum_{a=1}^{\infty} \sum_{b=1}^{2^a} \frac{1}{3^a} \\ &= \sum_{a=1}^{\infty} \left(\frac{2}{3}\right)^a < \infty. \end{aligned}$$

Define a process by $f(\omega) = u(\omega) + v(\omega)$ and

$$X_n(\omega) = f(T^n \omega).$$

Notice that $X_n \in \{0, 10^{-k}, \frac{2^{l_k}}{3^{a_k}}, k = 3, 4, \ldots\}$. Observe that $P(E_k) = 2^{-a_k}$ and

$$\sum_{k=3}^{\infty} P(E_k) = \sum_{a=1}^{\infty} \sum_{b=1}^{2^a} 2^{-a} = \sum_{a=1}^{\infty} 1 = \infty.$$

By the Borel–Cantelli lemma, a point ω belongs to E_k infinitely often. When $\omega \in E_k$,

$$T^{i_0} \omega \in D_k \qquad \text{for some } 0 \leq i_0 \leq 2^{l_k - a_k - 1} - 1.$$

For $\omega \in E_k$, we know that $X_{i_0}(\omega) = 10^{-k}$. At time $i_0 + 2^{l_k - a_k - 1} - 1$,

$$(T^{i_0 + 2^{l_k - a_k - 1} - 1}(\omega))_j = \begin{cases} 0, & \text{if } j = 1, \\ 1, & \text{if } 1 < j \leq l_k - 1, \\ \omega_j, & \text{otherwise.} \end{cases}$$

Let's compute for a fixed i_0 such that $T^{i_0}\omega \in D_k$ (i.e., $X_{i_0} = 10^{-k}$)

$$E(X_{i_0+2^{l_k-a_k}}|X_0^{i_0+2^{l_k-a_k}-1}).$$

Take $N = 2^{l_k-a_k}$ and consider

$$\frac{1}{N}\sum_{n=1}^{N}|E(X_n|X_0^{n-1}) - E(X_n|X_{-\infty}^{n-1})|.$$

For $\omega \in T^{-i_0}D_k$ (i.e., $X_{i_0} = 10^{-k}$), we know that

$$(T^{i_0+2^{l_k-a_k-1}}\omega)_j = \begin{cases} 1, & \text{if } 1 \le j \le l_k - 1, \\ \omega_j, & \text{otherwise.} \end{cases}$$

Therefore if $X_{i_0+2^{l_k-a_k-1}} > 0$, then we must have

$$T^{i_0+2^{l_k-a_k-1}}\omega \in C_k \cup \bigcup_{j>k}(C_j \cup D_j)$$

(because if $k' < k$ then $l_{k'} < l_k$ and $C_{k'}$, $D_{k'}$ are defined by zero values of ω_i with $i < l_k$) and

$$\begin{aligned}
&E(X_{i_0+2^{l_k-a_k-1}}|X_0^{i_0+2^{l_k-a_k-1}-1}) \\
&\quad= \frac{2^{l_k}/3^{a_k}2^{-l_k} + \sum_{j>k}2^{l_j}/3^{a_j}2^{-l_j} + \sum_{j>k}10^{-j}2^{-l_j} + 0}{P(D_k)} \\
&\quad\ge \frac{(2/3)^{a_k+1}}{2\cdot 2^{-l_k}} \\
&\quad= \frac{1}{2}2^{l_k}\left(\frac{2}{3}\right)^{a_k+1}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&E(X_{i_0+2^{l_k-a_k}}|X_0^{i_0+2^{l_k-a_k}-1}) \\
&\quad= \frac{2^{l_k}/3^{a_k}2^{-l_k} + \sum_{j>k}2^{l_j}/3^{a_j}2^{-l_j} + \sum_{j>k}10^{-j}2^{-l_j} + 0}{P(D_k)} \\
&\quad\le \frac{10^{-k-1} + \sum_{i=0}^{\infty}(2/3)^{a_k+i}}{2\cdot 2^{-l_k}} \\
&\quad= \frac{1}{2}2^{l_k}\left(10^{-k-1} + \left(\frac{2}{3}\right)^{a_k}3\right) \\
&\quad\le 4\cdot 2^{l_k}\left(\frac{2}{3}\right)^{a_k}.
\end{aligned}$$

On the other hand, $X_{-\infty}^{i_0+2^{l_k-a_k-1}-1}$ determines exactly the value of $X_{i_0+2^{l_k-a_k-1}}$. There are four cases. If $X_{i_0+2^{l_k-a_k-1}}$ is equal with 0, $\frac{2^{l_k}}{3^{a_k}}$, or for some $k<k':10^{-k'}$ or $\frac{2^{l_{k'}}}{3^{a_{k'}}}$. That is,

$$
\begin{aligned}
&E(X_{i_0+2^{l_k-a_k-1}}|X_{-\infty}^{i_0+2^{l_k-a_k-1}-1})\\
&\qquad=\begin{cases}
\dfrac{2^{l_k}}{3^{a_k}}, & \text{if } T^{i_0+2^{l_k-a_k-1}}\omega\in C_k,\\
10^{-k'}, & \text{if } T^{i_0+2^{l_k-a_k-1}}\omega\in D_{k'} \text{ for some } k<k',\\
\dfrac{2^{l_{k'}}}{3^{a_{k'}}}, & \text{if } T^{i_0+2^{l_k-a_k-1}}\omega\in C_{k'} \text{ for some } k<k',\\
0, & \text{if otherwise.}
\end{cases}
\end{aligned}
$$

Now

$$
\begin{aligned}
&|E(X_{i_0+2^{l_k-a_k-1}}|X_0^{i_0+2^{l_k-a_k-1}-1})-E(X_{i_0+2^{l_k-a_k-1}}|X_{-\infty}^{i_0+2^{l_k-a_k-1}-1})|\\
&\qquad\geq\begin{cases}
0.52^{l_k}\dfrac{2^{a_k}}{3^{a_k}}, & \text{if } T^{i_0+2^{l_k-a_k-1}}\omega\in C_k,\\
10^{-k'}, & \text{if } T^{i_0+2^{l_k-a_k-1}}\omega\in D_{k'} \text{ for some } k<k',\\
2^{l_k}, & \text{if } T^{i_0+2^{l_k-a_k-1}}\omega\in C_{k'} \text{ for some } k<k',\\
0.52^{l_k}\left(\dfrac{2}{3}\right)^{a_k+1}, & \text{if otherwise,}
\end{cases}
\end{aligned}
$$

where we assumed that $l_{k'}-2a_{k'}>l_k$ if $k'>k$. Now

$$
\begin{aligned}
&|E(X_{i_0+2^{l_k-a_k-1}}|X_{-\infty}^{i_0+2^{l_k-a_k-1}-1})-E(X_{i_0+2^{l_k-a_k-1}}|X_{-\infty}^{i_0+2^{l_k-a_k-1}-1})|\\
&\qquad\geq\frac{1}{4}2^{l_k}\left(\frac{2}{3}\right)^{a_k+1}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\frac{1}{N}\sum_{n=1}^{N}|E(X_n|X_0^{n-1})-E(X_n|X_{-\infty}^{n-1})| &\geq \frac{1}{N}\frac{1}{4}2^{l_k}\left(\frac{2}{3}\right)^{a_k+1}\\
&=2^{-l_k+a_k}\frac{1}{4}2^{l_k}\left(\frac{2}{3}\right)^{a_k+1}\\
&=\frac{1}{6}\left(\frac{4}{3}\right)^{a_k}.
\end{aligned}
$$

Since $\limsup_{k\to\infty}a_k=\infty$, the proof of Theorem 2 will be complete as soon as we verify that the process is ergodic and finitarily Markovian. The first property follows from the fact that T is an ergodic transformation. To see

the second, what we need to do is to show that the values of $f(T^{-n}\omega)$ will reveal to us more and more of the values of ω_m as n increases. Almost every point is in infintely many $T^{2^{l_j-a_j}}E_j$'s. For any such j, there is a unique $i<2^{l_j-a_j}$ such that $T^{i-2^{l_j-a_j}}\omega \in D_j$ and this is revealed to us by the value of f at the point in the negative orbit of ω. This information will give us the values of ω_m for all m up to l_j-a_j and this completes the proof. $\square$

Remark 2. The referee pointed out that a simpler and equivalent formulation of the first statement of the theorem above is as follows.

Let $\mathcal{X}=\{0,10^{-k},\frac{2^k}{3^m},k=1,2,\ldots,m=1,2,\ldots\}$. There exists a stationary and ergodic finitarily Markovian time series $\{X_n\}$ taking values from $\mathcal{X}$ such that $E|X_0|<\infty$ and

$$\limsup_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}|E(X_n|X_0^{n-1})|=\infty$$

almost surely.

[This is because $E(X_n|X_{-\infty}^{n-1})(\omega)=E(X_0|X_{-\infty}^{-1})(T^n\omega)$ and by the ergodic theorem

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}E(X_n|X_{-\infty}^{n-1})=E(X_0)<\infty$$

almost surely.]

Theorem 3. *Let $\mathcal{X}=\{0,10^{-k},\frac{2^k}{3^m}k=1,2,\ldots,m=1,2,\ldots\}$. Suppose $h_m:\mathcal{X}^m\to\mathbb{R}$ is a scheme that for any bounded ergodic finitarily Markovian process $\{Y_n\}$ taking values from $\mathcal{X}$, almost surely satisfies*

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}|E(Y_n|Y_0^{n-1})-h_n(Y_0^{n-1})|=0.$$

Then there is an ergodic finitarily Markovian process $\{X_n\}$ taking values from $\mathcal{X}$ for which

$$E|X_0|<\infty$$

and

$$\limsup_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}|E(X_n|X_0^{n-1})-h_n(X_0^{n-1})|=\infty$$

almost surely.

PROOF. I. *A Master process.* We shall prepare a master process with many possibilities for constructing a process such as in the earlier example with l_k in a fashion that will be dictated by the estimation scheme. For $1 \le j \le n$, define

$$q(n,j) = (n^2 + j)!$$

and sets

$$C_{q(n,j)} = \{\omega : \omega_i = 1 \text{ for } 1 \le i < q(n,j), \omega_{q(n,j)} = 0\},$$

clearly $P(C_{q(n,j)}) = 2^{-q(n,j)}$. Let

$$D_{q(n,j)} = \{\omega : \omega_i = 1 \text{ for } 1 \le i < q(n,j) - j, \\ \omega_{q(n,j)-j} = 0, \omega_i = 1 \text{ for } q(n,j) - j < i < q(n,j)\}$$

and

$$E_{q(n,j)} = \bigcup_{i=0}^{2^{q(n,j)-j-1}-1} T^{-i} D_{q(n,j)}.$$

Notice that

$$E_{q(n,j)} = \{\omega : \omega_{q(n,j)-j} = 0, \omega_i = 1, \text{ for all } q(n,j) - j < i < q(n,j)\}$$

and it follows that the sets $\{E_{q(n,j)}, 1 \le j \le n, n \in \mathbf{N}\}$ are mutually independent. Letting

$$u(\omega) = \sum_{n=1}^{\infty} \sum_{j=1}^{n} 10^{-q(n,j)} I_{D_{q(n,j)}}(\omega)$$

the master process is defined by $Y_n(\omega) = u(T^n \omega)$. For later use, observe that the $D_{q(n,j)}$'s are disjoint.

We will need the following easy consequence of our assumption on the estimators h_n, namely that for any bounded process Y_n defined on Ω as in the theorem and for any k there is an integer N_k and a set $H_k \subset \Omega$ with $P(H_k) \ge 1 - 2^{-k}$ and for all $\omega \in H_k$ and $m \ge N_k$ we have: $|h_m(Y_0, \ldots, Y^{(k-1)})| \le \frac{m}{10}$.

II. *The construction.* We shall now define a sequence l_k, $k = 2^{a_k} + b_k$, $1 \le b \le 2^a$ inductively, together with functions v_k which are bounded. As k tends to infinity, the v_k will converge to v and we will use $u + v$ to get our desired process. We may take $v_2 = 0$ to start the inductive construction.

Assume that we have already defined $l_3 < l_4 < \cdots < l_{k-1}$ a subsequence of the $q(n,j)$'s and

$$v_{k-1} = \sum_{i=3}^{k-1} \left(\frac{2^{l_i}}{3^{a_i}} \right) I_{C_{l_i}}$$

we want to define l_k and v_k. Recalling the notation $k-1 = 2^{a_{k-1}} + b_{k-1}$, we have that $b_{k-1} = b_k - 1$ unless $k-1 = 2^a$, in which case $a_{k-1} = a-1$ and $b_{k-1} = 2^{a-1}$.

Since v_{k-1} is bounded, the process defined by

$$X_n^{(k-1)} = f_{k-1}(T^n\omega) = u(T^n\omega) + v_{k-1}(T^n\omega)$$

is bounded. Now, by assumption, there is an N_k and a set H_k with $P(H_k) \geq 1 - 2^{-k}$ and for all $\omega \in H_k$ and $m \geq N_k$ we know that

$$|h_m(X_0^{(k-1)}, \ldots, X_{m-1}^{(k-1)})| \leq \frac{m}{10}.$$

Choose n large enough so that $2^{q(n,a_k)-a_k} > 10N_k$ and we make sure that $q(n,a_k) - a_k > 10l_{k-1}$. Set

$$l_k = q(n, a_k)$$

and

$$v_k = v_{k-1} + \left(\frac{2^{l_k}}{3^{a_k}}\right) I_{C_{l_k}}.$$

This defines a new process

$$X_n^{(k)}(\omega) = f_k(T^n\omega) = u(T^n\omega) + v_k(T^n\omega).$$

It is important to observe that if for some $i_0 \leq 2^{l_k - a_k - 1}$ we have $T^{i_0}\omega \in D_{l_k}$ then for all $0 \leq j \leq i_0 + 2^{l_k - a_k - 1} - 1$

$$X_j^{(k)}(\omega) = X_j^{(k-1)}(\omega).$$

This is because the way C_{l_k} is defined, we know that $T^{i_0 + 2^{l_k - a_k - 1}}\omega$ can be in C_{l_k} which implies that earlier iterates of ω cannot be there. Indeed,

$$C_{q(n,j)} \subset T^{2^{l_k - a_k - 1}} D_{l_k} \qquad \text{for all } q(n,j) \geq l_k,$$

which implies that during all the later stages of the construction the values of $X_i^{(k-1)}$ in this range will not change. So we will have for

$$v = \sum_{k=3}^{\infty} \left(\frac{2^{l_k}}{3^{a_k}}\right) I_{C_{l_k}}$$

and

$$X_n(\omega) = f(T^n\omega) = u(T^n\omega) + v(T^n\omega)$$

that

$$X_j(\omega) = X_j^{(k-1)}(\omega) \qquad \text{for all } 0 \leq j \leq i_0 + 2^{l_k - a_k - 1} - 1,$$

if $T^{i_0}\omega \in D_{l_k}$.

It is clear that if the l_k's are chosen large enough so that for all $k' > k$ $l_k < l_{k'} - 2a_{k'}$:

- the sets $\{C_k, D_k\}_{k=3}^{\infty}$ are disjoint,
- the intervals $[l_k - a_k, l_k - 1]$ are also disjoint and therefore the sets E_{l_k} are independent.

The signaling function u is bounded and the main contributor to f will be

$$v(\omega) = \sum_{k=3}^{\infty} \frac{2^{l_k}}{3^{a_k}} I_{C_{l_k}}(\omega).$$

Clearly,

$$E(v(\omega)) = \sum_{k=3}^{\infty} \frac{2^{l_k}}{3^{a_k}} P(C_{l_k}) = \sum_{k=3}^{\infty} \frac{1}{3^{a_k}} = \sum_{a=1}^{\infty} \sum_{b=1}^{2^a} \frac{1}{3^a} = \sum_{a=1}^{\infty} \left(\frac{2}{3}\right)^a < \infty.$$

Define a process by $f(\omega) = u(\omega) + v(\omega)$ and

$$X_n(\omega) = f(T^n \omega).$$

Note that $X_n \in \mathcal{X}$ as advertised.

III. *Checking the properties*. Observe that $P(E_{l_k}) = 2^{-a_k}$ and

$$\sum_{k=3}^{\infty} P(E_{l_k}) = \sum_{a=1}^{\infty} \sum_{b=1}^{2^a} 2^{-a} = \sum_{a=1}^{\infty} 1 = \infty.$$

By the Borel–Cantelli lemma, a point ω belongs to E_{l_k} infinitely often. In addition, since $P(H_k) > 1 - 2^{-k}$, almost every point will belong to H_k for all sufficiently large k. Suppose then that $\omega \in E_{l_k} \cap H_k$. When $\omega \in E_{l_k}$,

$$T^{i_0}\omega \in D_k \qquad \text{for some } 0 \le i_0 \le 2^{l_k - a_k - 1} - 1.$$

For $\omega \in E_{l_k}$, we know that $X_{i_0}(\omega) = 10^{-l_k}$. At time $i_0 + 2^{l_k - a_k - 1} - 1$,

$$(T^{i_0 + 2^{l_k - a_k - 1} - 1}(\omega))_j = \begin{cases} 0, & \text{if } j = 1, \\ 1, & \text{if } 1 < j \le l_k - 1, \\ \omega_j, & \text{otherwise.} \end{cases}$$

Let's compute for a fixed i_0 such that $T^{i_0}\omega \in D_{l_k}$ (i.e., $X_{i_0} = 10^{-l_k}$)

$$E(X_{i_0 + 2^{l_k - a_k}} | X_0^{i_0 + 2^{l_k - a_k - 1}}).$$

For $\omega \in T^{-i_0} D_{l_k}$ (i.e., $X_{i_0} = 10^{-l_k}$) we know that

$$(T^{i_0 + 2^{l_k - a_k - 1}} \omega)_j = \begin{cases} 1, & \text{if } 1 \le j \le l_k - 1, \\ \omega_j, & \text{otherwise.} \end{cases}$$

Therefore if $X_{i_0 + 2^{l_k - a_k - 1}} > 0$, then we must have

$$T^{i_0 + 2^{l_k - a_k - 1}} \omega \in C_{l_k} \cup \bigcup_{m > k} C_m \cup \bigcup_{1 \le n, 1 \le j \le 2^n \,:\, q(n,j) > l_k} D_{q(n,j)},$$

because if $k' < k$ then $l_{k'} < l_k$ and the $C_{k'}$, are defined by zero values of ω_i with $i < l_k$, and similarly for $D_{q(n,j)}$ with $q(n,j) < l_k$,

$$\begin{aligned}
&E(X_{i_0+2^{l_k-a_k-1}} | X_0^{i_0+2^{l_k-a_k-1}-1}) \\
&\quad = \frac{2^{l_k}/3^{a_k} 2^{-l_k} + \sum_{j>k} 2^{l_j}/3^{a_j} 2^{-l_j} + \sum_{q(n,j)>l_k} 10^{-q(n,j)} 2^{-q(n,j)} + 0}{P(D_{l_k})} \\
&\quad \geq \frac{(2/3)^{a_k+1}}{2 \cdot 2^{-l_k}} \\
&\quad = \frac{1}{2} 2^{l_k} \left(\frac{2}{3}\right)^{a_k+1}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&E(X_{i_0+2^{l_k-a_k}} | X_0^{i_0+2^{l_k-a_k}-1}) \\
&\quad = \frac{2^{l_k}/3^{a_k} 2^{-l_k} + \sum_{j>k} 2^{l_j}/3^{a_j} 2^{-l_j} + \sum_{q(n,j)>l_k} 10^{-q(n,j)} 2^{-q(n,j)} + 0}{P(D_{l_k})} \\
&\quad \leq \frac{10^{-l_k} + \sum_{i=0}^{\infty} (2/3)^{a_k+i}}{2 \cdot 2^{-l_k}} \\
&\quad = \frac{1}{2} 2^{l_k} \left(10^{-k-1} + \left(\frac{2}{3}\right)^{a_k} 3\right) \\
&\quad \leq 4 \cdot 2^{l_k} \left(\frac{2}{3}\right)^{a_k}.
\end{aligned}$$

On the other hand, because $\omega \in H_k$ and our remark about $X_j = X_j^{(k-1)}$ for $0 \leq j \leq 2^{l_k-a_k-1} - 1$, we have that

$$|h_{i_0+2^{l_k-a_k-1}}(X_0^{i_0+2^{l_k-a_k-1}-1})| \leq \frac{i_0 + 2^{l_k-a_k-1} - 1}{10}.$$

Therefore, if we take $N = 2^{l_k-a_k}$

$$\begin{aligned}
\frac{1}{N} \sum_{n=1}^{N} |E(X_n | X_0^{n-1}) - h_n(X_0^{n-1})| &\geq \frac{1}{N} \frac{1}{4} 2^{l_k} \left(\frac{2}{3}\right)^{a_k+1} \\
&= 2^{-l_k+a_k} \frac{1}{4} 2^{l_k} \left(\frac{2}{3}\right)^{a_k+1} \\
&= \frac{1}{6} \left(\frac{4}{3}\right)^{a_k}.
\end{aligned}$$

Since $\limsup_{k\to\infty} a_k = \infty$, the proof of Theorem 3 is complete. $\square$

APPENDIX

The next result is a generalization of a result due to Elton; cf. Theorems 2 and 4 in Elton [7].

PROPOSITION 1. *For $n = 0, 1, 2, \dots$, let $\mathcal{F}_n$ be an increasing sequence of σ-fields, and X_n random variables measurable with respect to $\mathcal{F}_n$, be identically distributed with $E(|X_0| \log^+(|X_0|)) < \infty$. Let $g_n(X_n)$ be quantizing functions so that for all n, $|g_n(X_n) - X_n| \le 1$, and for an increasing sequence of sub σ-fields, $\mathcal{G}_n \subseteq \mathcal{F}_n$ such that $g_n(X_n) = Y_n$ is measurable with respect to $\mathcal{G}_n$, form the sequence of martingale differences*

$$Z_n = g_n(X_n) - E(g_n(X_n)|\mathcal{G}_{n-1}) = Y_n - E(Y_n|\mathcal{G}_{n-1}).$$

Then

$$E\left(\sup_{1 \le n} \left|\frac{1}{n}\sum_{i=1}^{n} Z_i\right|\right) < \infty \tag{A.1}$$

and

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} Z_i = 0 \qquad \textit{almost surely.} \tag{A.2}$$

PROOF. We follow Elton [7], who gave the proof when the martingale differences Z_n are identically distributed. Write

$$Y_n = Y_n' + Y_n'',$$

where $|Y_n'| \le n$ and $|Y_n''| > n$. Now

$$Z_n = Y_n' - E(Y_n'|\mathcal{G}_{n-1}) + Y_n'' - E(Y_n''|\mathcal{G}_{n-1}).$$

Since for any sequence of real numbers $\{a_i\}$,

$$\sup_{1 \le n} \frac{1}{n}\left|\sum_{i=1}^{n} a_i\right| \le 2\left(\sup_{1 \le n}\left|\sum_{i=1}^{n} \frac{1}{i} a_i\right|\right),$$

(cf. Lemma 7 in Elton [7]), letting

$$d_n = Y_n' - E(Y_n'|\mathcal{G}_{n-1})$$

and

$$e_n = Y_n'' - E(Y_n''|\mathcal{G}_{n-1})$$

we get

$$E\left(\sup_{1\le n}\frac{1}{n}\left|\sum_{i=1}^{n} Z_i\right|\right) \tag{A.3}$$

$$\le 2E\left(\sup_{1\le n}\left|\sum_{i=1}^{n}\frac{1}{i} Z_i\right|\right) \tag{A.4}$$

$$\le 2E\left(\sup_{1\le n}\left|\sum_{i=1}^{n}\frac{1}{i} d_i\right|\right) \tag{A.5}$$

$$+ 2E\left(\sup_{1\le n}\left|\sum_{i=1}^{n}\frac{1}{i} e_i\right|\right). \tag{A.6}$$

For (A.5) by Davis' inequality (valid for all martingale differences cf. e.g., Shiryayev [32], page 470), we get

$$2E\left(\sup_{1\le n}\left|\sum_{i=1}^{n}\frac{1}{i} d_i\right|\right) \le 2BE\left[\left(\sum_{i=1}^{\infty}\frac{1}{i^2}(d_i)^2\right)^{0.5}\right] \le 2B\left[E\left(\sum_{i=1}^{\infty}\frac{1}{i^2}(d_i)^2\right)\right]^{0.5}.$$

Now, $E((d_i)^2) \le E((Y_i')^2)$. But since $|Y_i - X_i| \le 1$, we get

$$E((Y_i')^2) = E((Y_i)^2 I_{\{|Y_i|\le i\}}) \le E((X_i+1)^2 I_{\{|X_i-1|\le i\}})$$

and the X_i's are identically distributed therefore

$$\begin{aligned}
&\sum_{i=1}^{\infty}\frac{1}{i^2}E((X_i+1)^2 I_{\{|X_i-1|\le i\}})\\
&\qquad = \sum_{i=1}^{\infty}\left(E((X_i+1)^2 I_{\{i-1<|X_i-1|\le i\}})\left(\sum_{j=i}^{\infty}\frac{1}{j^2}\right)\right)\\
&\qquad \le KE(|X_0|),
\end{aligned}$$

where K is a suitable constant (cf. the last line of the proof of Lemma 1 in Elton [7]).

For (A.6),

$$E|e_n| \le 2E|Y_n''| \le 2E((1+|X_n|)I_{\{|X_n|>n-1\}})$$

and now X_n' are identically distributed. Now since $E(|X|\log^+(|X|)) < \infty$, Lemma 2 in Elton [7] implies that

$$\sum_{n=1}^{\infty}\frac{1}{n}E((1+|X_n|)I_{\{|X_n|>n-1\}}) < \infty$$

and so

$$2E\left(\sup_{1\le n}\left|\sum_{i=1}^{n}\frac{1}{i}e_i\right|\right)\le 2\sum_{i=1}^{n}\frac{1}{i}E|e_i|$$
$$<\infty$$

and this completes the proof of (A.1).

Now, we prove (A.2). By (A.4),

$$U_n=\sum_{i=1}^{n}\frac{1}{i}Z_i$$

is a martingale with

$$\sup_{1\le n}E(|U_n|)<\infty$$

and by Doob's convergence theorem U_n converges almost surely. Then by Kronecker's lemma (cf. Shiryayev [32], page 365),

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}Z_i=0$$

almost surely. The proof of Proposition 1 is complete. □

PROPOSITION 2. *Let* $\{\phi_n,\mathcal{F}_n\}$ *be a martingale difference sequence. If, for some* $1<p<\infty$, $\sup_{1\le n}E(|\phi_n|^p)<\infty$ *then*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\phi_i=0 \qquad \textit{almost surely} \tag{A.7}$$

and

$$E\left(\sup_{1\le n}\left|\frac{1}{n}\sum_{i=1}^{n}\phi_i\right|^p\right)<\infty. \tag{A.8}$$

PROOF. Choose a positive integer K such that $K(p-1)>1$. Define

$$f_n=\frac{1}{n}\sum_{i=1}^{n}\phi_i.$$

Assume first that $1<p\le 2$. Now by Theorem 2 in von Bahr and Esseen [33],

$$E(|f_n|^p)\le 2\frac{n}{n^p}\sup_{1\le i}E(|\phi_i|^p)=2\frac{\sup_{1\le i}E(|\phi_i|^p)}{n^{(p-1)}}. \tag{A.9}$$

Define

$$F=\sum_{n=1}^{\infty}|f_{n^K}|^p.$$

By (A.9), and since by assumption $\sup_{1\le n} E(|\phi_n|^p)<\infty$, $K(p-1)>1$,

$$E(F)=2\sum_{n=1}^{\infty}\frac{\sup_{1\le i} E(|\phi_i|^p)}{n^{K(p-1)}}<\infty. \tag{A.10}$$

Define

$$g_n=\max_{1\le k<(n+1)^K-n^K}|f_{n^K}-f_{n^K+k}|^p$$

and let

$$G=\sum_{n=1}^{\infty}g_n.$$

To complete the proof of (A.7) and (A.8), it is enough to show that $E(F+G)<\infty$. By (A.10), it is enough to show that $E(G)<\infty$. Now for some $m=n^K+k$, $1\le k<(n+1)^K-n^K$,

$$f_m=(f_{n^K+k}-f_{n^K})+f_{n^K}$$

and

$$|f_m|^p\le 2^p(|f_{n^K+k}-f_{n^K}|^p+|f_{n^K}|^p)\le 2^p(g_n+|f_{n^K}|^p)\le 2^p(G+F).$$

Now

$$|f_{n^K+k}-f_{n^K}|=\left(\frac{1}{n^K}-\frac{1}{n^K+k}\right)\sum_{i=1}^{n^K}\phi_i-\frac{1}{n^K+k}\sum_{j=1}^{k}\phi_{n^K+j}$$

and so

$$\begin{aligned}|f_{n^K+k}-f_{n^K}|^p&\le 2^p\left(\left|\frac{k}{n^K(n^K+k)}\sum_{i=1}^{n^K}\phi_i\right|^p+\left|\frac{1}{n^K+k}\sum_{j=1}^{k}\phi_{n^K+j}\right|^p\right)\\&\le 2^p\left(\left|\frac{(n+1)^K-n^K-1}{n^Kn^K}\sum_{i=1}^{n^K}\phi_i\right|^p+\left|\frac{1}{n^K}\sum_{j=1}^{k}\phi_{n^K+j}\right|^p\right).\end{aligned}$$

Now

$$g_n\le 2^p\left(\left|\frac{(n+1)^K-n^K}{n^Kn^K}\sum_{i=1}^{n^K}\phi_i\right|^p+\left|\frac{1}{n^{Kp}}\max_{1\le k<(n+1)^K-n^K}\sum_{j=1}^{k}\phi_{n^K+j}\right|^p\right).$$

Now by von Bahr and Eseen [33] and Doob's inequality (cf. e.g., Theorem 1, Chapter 3 in Shiryayev [32]),

$$\begin{aligned} E(g_n) &\le 2^p\left(\frac{(n+1)^K-n^K}{n^{2K}}\right)^p 2n^K\left(\sup_{1\le i}E(|\phi_i|^p)\right) \\ &\qquad + \left(\frac{p}{(p-1)}\right)^p\frac{1}{n^K}E\left(\left|\sum_{j=1}^{(n+1)^K-n^K}\phi_{n^K+j}\right|^p\right) \\ &\le 2^p\left(\frac{(n+1)^K-n^K}{n^{2K}}\right)^p 2n^K\left(\sup_{1\le i}E(|\phi_i|^p)\right) \\ &\qquad + \left(\frac{p}{(p-1)}\right)^p\left(\frac{(n+1)^K-n^K}{n^K}\right)^p\sup_{1\le i}E(|\phi_i|^p) \end{aligned}$$

and the right-hand side is summable. We have completed the proof for $1 < p \le 2$.

Now assume $2 < p < \infty$. By the theorem of Dharmadhikari, Fabian and Jogdeo [6],

$$E(|f_n|^p) \le C(p)\frac{\sup_{1\le i}E(|\phi_i|^p)}{n^{p/2}}.$$

Applying this one gets that

$$E\left(\sum_{n=1}^{\infty}|f_n|^p\right) \le \sum_{n=1}^{\infty}C(p)\frac{\sup_{1\le i}E(|\phi_i|^p)}{n^{p/2}} < \infty.$$

Thus,

$$\sum_{n=1}^{\infty}|f_n|^p < \infty \qquad \text{almost surely}$$

and this yields (A.7) and (A.8). The proof of Proposition 2 is complete. □

REMARK 3. The referee pointed out that the second statement of the preposition above could be proved in a simpler way as follows. By maximal Doob inequality and Burkholder inequality, we obtain

$$\left(E\sup_n\left|\frac{1}{n}\sum_{i=1}^{n}\phi_i\right|^p\right)^{1/p} \le 2p\max\left\{1, \frac{1}{(p-1)^2}\right\}\left[E\left(\sum_{i=1}^{\infty}\left(\frac{\phi_i}{i}\right)^2\right)^{p/2}\right]^{1/p}.$$

Now if $p \ge 2$, then by the triangle inequality in $L_{p/2}$,

$$\left\{\left[E\left(\sum_{i=1}^{\infty}\left(\frac{\phi_i}{i}\right)^2\right)^{p/2}\right]^{2/p}\right\}^{1/2} \le \left[\sum_{i=1}^{\infty}\left(E\frac{|\phi_i|^p}{i^p}\right)^{2/p}\right]^{1/2}$$

$$\leq \left(\sum_{i=1}^{\infty}\frac{1}{i^p}\right)^{1/2}\sup_i(E|\phi_i|^p)^{1/p}.$$

If $p \leq 2$, then since

$$\left(\sum_i a_i\right)^{p/2} \leq \sum_i (a_i)^{p/2}$$

for all positive numbers a_i we get

$$\left[E\left(\sum_{i=1}^{\infty}\left(\frac{\phi_i}{i}\right)^2\right)^{p/2}\right]^{1/p} \leq \sum_{i=1}^{\infty}\left(E\frac{|\phi_i|^p}{i^p}\right)^{1/p} \leq \left(\sum_{i=1}^{\infty}\frac{1}{i^2}\right)^{1/2}\sup_i(E|\phi_i|^p)^{1/p}.$$

Thus, in each case it is

$$\left(E\sup_n\left|\frac{1}{n}\sum_{i=1}^{n}\phi_i\right|^p\right)^{1/p} \leq C_p\sup_i(E|\phi_i|^p)^{1/p}.$$

REFERENCES

[1] Algoet, P. (1992). Universal schemes for prediction, gambling and portfolio selection. *Ann. Probab.* **20** 901–941. MR1159579

[2] Algoet, P. H. (1994). The strong law of large numbers for sequential decisions under uncertainty. *IEEE Trans. Inform. Theory* **40** 609–633. MR1295308

[3] Algoet, P. (1999). Universal schemes for learning the best nonlinear predictor given the infinite past and side information. *IEEE Trans. Inform. Theory* **45** 1165–1185. MR1686250

[4] Aaronson, J. (1997). *An Introduction to Infinite Ergodic Theory. Mathematical Surveys and Monographs* **50**. Amer. Math. Soc., Providence, RI. MR1450400

[5] Bailey, D. H. (1976). Sequential schemes for classifying and predicting ergodic processes. Ph.D. thesis, Stanford Univ.

[6] Dharmadhikari, S. W., Fabian, V. and Jogdeo, K. (1968). Bounds on the moments of martingales. *Ann. Math. Statist.* **39** 1719–1723. MR0230363

[7] Elton, J. (1981). A law of large numbers for identically distributed martingale differences. *Ann. Probab.* **9** 405–412. MR614626

[8] Feder, M. and Merhav, N. (1998). Universal prediction. *IEEE Trans. Inform. Theory* **44** 2124–2147. MR1658815

[9] Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002). *A Distribution-free Theory of Nonparametric Regression*. Springer, New York. MR1920390

[10] Györfi, L., Morvai, G. and Yakowitz, S. J. (1998). Limits to consistent on-line forecasting for ergodic time series. *IEEE Trans. Inform. Theory* **44** 886–892. MR1607704

[11] Györfi, L., Lugosi, G. and Morvai, G. (1999). A simple randomized algorithm for sequential prediction of ergodic time series. *IEEE Trans. Inform. Theory* **45** 2642–2650. MR1725166

[12] Györfi, L. and Ottucsák, G. (2007). Sequential prediction of unbounded stationary time series. *IEEE Trans. Inform. Theory* **53** 1866–1872. MR2317147

[13] Kalikow, S., Katznelson, Y. and Weiss, B. (1992). Finitarily deterministic generators for zero entropy systems. *Israel J. Math.* **79** 33–45. MR1195252

[14] Krengel, U. (1985). *Ergodic Theorems. de Gruyter Studies in Mathematics* **6**. de Gruyter, Berlin. MR797411

[15] Maker, P. T. (1940). The ergodic theorem for a sequence of functions. *Duke Math. J.* **6** 27–30. MR0002028

[16] Morvai, G. (1994). Estimation of conditional distribution for stationary time series. Ph.D. thesis, Technical Univ. Budapest.

[17] Morvai, G., Yakowitz, S. J. and Algoet, P. (1997). Weakly convergent nonparametric forecasting of stationary time series. *IEEE Trans. Inform. Theory* **43** 483–498. MR1447529

[18] Morvai, G., Yakowitz, S. and Györfi, L. (1996). Nonparametric inference for ergodic, stationary time series. *Ann. Statist.* **24** 370–379. MR1389896

[19] Morvai, G. (2003). Guessing the output of a stationary binary time series. In *Foundations of Statistical Inference (Shoresh, 2000)* 207–215. Physica, Heidelberg. MR2017826

[20] Morvai, G. and Weiss, B. (2003). Forecasting for stationary binary time series. In *Proceedings of the Eighth Vilnius Conference on Probability Theory and Mathematical Statistics, Part II (2002)* **79** 25–34. MR2021874

[21] Morvai, G. and Weiss, B. (2005). Inferring the conditional mean. *Theory Stoch. Process.* **11** 112–120. MR2327452

[22] Morvai, G. and Weiss, B. (2005). On classifying processes. *Bernoulli* **11** 523–532. MR2146893

[23] Morvai, G. and Weiss, B. (2005). Limitations on intermittent forecasting. *Statist. Probab. Lett.* **72** 285–290. MR2153125

[24] Morvai, G. and Weiss, B. (2005). Prediction for discrete time series. *Probab. Theory Related Fields* **132** 1–12. MR2136864

[25] Morvai, G. and Weiss, B. (2004). Intermittent estimation of stationary time series. *Test* **13** 525–542. MR2154012

[26] Morvai, G. and Weiss, B. (2007). On estimating the memory for finitarily Markovian processes. *Ann. Inst. H. Poincaré Probab. Statist.* **43** 15–30. MR2288267

[27] Morvai, G. and Weiss, B. (2005). Forward estimation for ergodic time series. *Ann. Inst. H. Poincaré Probab. Statist.* **41** 859–870. MR2165254

[28] Nobel, A. B. (2003). On optimal sequential prediction for general processes. *IEEE Trans. Inform. Theory* **49** 83–98. MR1965889

[29] Ornstein, D. S. (1978). Guessing the next output of a stationary process. *Israel J. Math.* **30** 292–296. MR508271

[30] Ryabko, B. Y. (1988). Prediction of random sequences and universal coding. *Probl. Inf. Transm.* **24** 3–14. MR955983

[31] Shields, P. C. (1996). *The Ergodic Theory of Discrete Sample Paths. Graduate Studies in Mathematics* **13**. Amer. Math. Soc., Providence, RI. MR1400225

[32] Shiryayev, A. N. (1984). *Probability. Graduate Texts in Mathematics* **95**. Springer, New York. MR737192

[33] von Bahr, B. and Esseen, C. G. (1965). Inequalities for the rth absolute moment of a sum of random variables, $1 \le r \le 2$. *Ann. Math. Statist.* **36** 299–303.

[34] Weiss, B. (2000). *Single Orbit Dynamics. CBMS Regional Conference Series in Mathematics* **95**. Amer. Math. Soc., Providence, RI. MR1727510

MTA-BME Stochastics Research Group
Institute of Mathematics
Egry József utca 1
Building H
Budapest, 1111
Hungary
E-mail: morvai@math.bme.hu
URL: http://www.math.bme.hu/~morvai/

Hebrew University of Jerusalem
Institute of Mathematics
Jerusalem 91904
Israel
E-mail: weiss@math.huji.ac.il